A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts

Bo Pang and Lillian Lee
Department of Computer Science
Cornell University
Ithaca, NY 14853-7501
{pabo,llee}@cs.cornell.edu

Abstract

Sentiment analysis seeks to identify the viewpoint(s) underlying a text span; an example application is classifying a movie review as “thumbs up” or “thumbs down”. To determine this sentiment polarity, we propose a novel machine-learning method that applies text-categorization techniques to just the subjective portions of the document. Extracting these portions can be implemented using efficient techniques for finding minimum cuts in graphs; this greatly facilitates incorporation of cross-sentence contextual constraints.


1 Introduction

The computational treatment of opinion, sentiment, and subjectivity has recently attracted a great deal of attention (see references), in part because of its potential applications. For instance, information-extraction and question-answering systems could flag statements and queries regarding opinions rather than facts (Cardie et al., 2003). Also, it has proven useful for companies, recommender systems, and editorial sites to create summaries of people’s experiences and opinions that consist of subjective expressions extracted from reviews (as is commonly done in movie ads) or even just a review’s polarity — positive (“thumbs up”) or negative (“thumbs down”).

Document polarity classification poses a significant challenge to data-driven methods, resisting traditional text-categorization techniques (Pang, Lee, and Vaithyanathan, 2002). Previous approaches focused on selecting indicative lexical features (e.g., the word “good”), classifying a document according to the number of such features that occur anywhere within it. In contrast, we propose the following process: (1) label the sentences in the document as either subjective or objective, discarding the latter; and then (2) apply a standard machine-learning classifier to the resulting extract. This can prevent the polarity classifier from considering irrelevant or even potentially misleading text: for example, although the sentence “The protagonist tries to protect her good name” contains the word “good”, it tells us nothing about the author’s opinion and in fact could well be embedded in a negative movie review. Also, as mentioned above, subjectivity extracts can be provided to users as a summary of the sentiment-oriented content of the document.

Our results show that the subjectivity extracts we create accurately represent the sentiment information of the originating documents in a much more compact form: depending on choice of downstream polarity classifier, we can achieve highly statistically significant improvement (from 82.8% to 86.4%) or maintain the same level of performance for the polarity classification task while retaining only 60% of the reviews’ words. Also, we explore extraction methods based on a minimum cut formulation, which provides an efficient, intuitive, and effective means for integrating inter-sentence-level contextual information with traditional bag-of-words features.

2 Method

2.1 Architecture

One can consider document-level polarity classification to be just a special (more difficult) case of text categorization with sentiment- rather than topic-based categories. Hence, standard machine-learning classification techniques, such as support vector machines (SVMs), can be applied to the entire documents themselves, as was done by Pang, Lee, and Vaithyanathan (2002). We refer to such classification techniques as default polarity classifiers.

However, as noted above, we may be able to im-
prove polarity classification by removing objective sentences (such as plot summaries in a movie review). We therefore propose, as depicted in Figure 1, to first employ a subjectivity detector that determines whether each sentence is subjective or not: discarding the objective ones creates an extract that should better represent a review’s subjective content to a default polarity classifier.

To our knowledge, previous work has not integrated sentence-level subjectivity detection with document-level sentiment polarity. Yu and Hatzivassiloglou (2003) provide methods for sentence-level analysis and for determining whether a document is subjective or not, but do not combine these two types of algorithms or consider document polarity classification. The motivation behind the single-sentence selection method of Beineke et al. (2004) is to reveal a document’s sentiment polarity, but they do not evaluate the polarity-classification accuracy that results.

2.2 Context and Subjectivity Detection

As with document-level polarity classification, we could perform subjectivity detection on individual sentences by applying a standard classification algorithm on each sentence in isolation. However, modeling proximity relationships between sentences would enable us to leverage coherence: text spans occurring near each other (within discourse boundaries) may share the same subjectivity status, other things being equal (Wiebe, 1994).

We would therefore like to supply our algorithms with pair-wise interaction information, e.g., to specify that two particular sentences should ideally receive the same subjectivity label but not state which label this should be. Incorporating such information is somewhat unnatural for classifiers whose input consists simply of individual feature vectors, such as Naive Bayes or SVMs, precisely because such classifiers label each test item in isolation. One could define synthetic features or feature vec-

tors to attempt to overcome this obstacle. However, we propose an alternative that avoids the need for such feature engineering: we use an efficient and intuitive graph-based formulation relying on finding minimum cuts. Our approach is inspired by Blum and Chawla (2001), although they focused on similarity between items (the motivation being to combine labeled and unlabeled data), whereas we are concerned with physical proximity between the items to be classified; indeed, in computer vision, modeling proximity information via graph cuts has led to very effective classification (Boykov, Veksler, and Zabih, 1999).

2.3 Cut-based classification

Figure 2 shows a worked example of the concepts in this section.

Suppose we have \( n \) items \( x_1, \ldots, x_n \) to divide into two classes \( C_1 \) and \( C_2 \), and we have access to two types of information:

- **Individual** scores \( \text{ind}_j(x_i) \): non-negative estimates of each \( x_i \)'s preference for being in \( C_j \) based on just the features of \( x_i \) alone; and
- **Association** scores \( \text{assoc}(x_i, x_k) \): non-negative estimates of how important it is that \( x_i \) and \( x_k \) be in the same class.\(^1\)

We would like to maximize each item’s “net happiness”: its individual score for the class it is assigned to, minus its individual score for the other class. But, we also want to penalize putting tightly-associated items into different classes. Thus, after some algebra, we arrive at the following optimization problem: assign the \( x_i \)s to \( C_1 \) and \( C_2 \) so as to minimize the partition cost

\[
\sum_{x \in C_1} \text{ind}_2(x) + \sum_{x \in C_2} \text{ind}_1(x) + \sum_{x_i \in C_1, x_k \in C_2} \text{assoc}(x_i, x_k).
\]

The problem appears intractable, since there are \( 2^n \) possible binary partitions of the \( x_i \)’s. However, suppose we represent the situation in the following manner. Build an undirected graph \( G \) with vertices \( \{v_1, \ldots, v_n, s, t\} \); the last two are, respectively, the source and sink. Add \( n \) edges \( (s, v_i) \), each with weight \( \text{ind}_1(x_i) \), and \( n \) edges \( (v_i, t) \), each with weight \( \text{ind}_2(x_i) \). Finally, add \( \binom{n}{2} \) edges \( (v_i, v_k) \), each with weight \( \text{assoc}(x_i, x_k) \). Then, cuts in \( G \) are defined as follows:

**Definition 1** A cut \( (S, T) \) of \( G \) is a partition of its nodes into sets \( S = \{s\} \cup S' \) and \( T = \{t\} \cup T' \), where \( s \not\in S', t \not\in T' \). Its cost \( \text{cost}(S, T) \) is the sum

\(^1\)Asymmetry is allowed, but we used symmetric scores.
of the weights of all edges crossing from \( S \) to \( T \). A minimum cut of \( G \) is one of minimum cost.

Observe that every cut corresponds to a partition of the items and has cost equal to the partition cost. Thus, our optimization problem reduces to finding minimum cuts.

**Practical advantages** As we have noted, formulating our subjectivity-detection problem in terms of graphs allows us to model item-specific and pairwise information independently. Note that this is a very flexible paradigm. For instance, it is perfectly legitimate to use knowledge-rich algorithms employing deep linguistic knowledge about sentiment indicators to derive the individual scores. And we could also simultaneously use knowledge-lean methods to assign the association scores. Interestingly, Yu and Hatzivassiloglou (2003) compared an individual-preference classifier against a relationship-based method, but didn’t combine the two; the ability to coordinate such algorithms is precisely one of the strengths of our approach.

But a crucial advantage specific to the utilization of a minimum-cut-based approach is that we can use maximum-flow algorithms with polynomial asymptotic running times — and near-linear running times in practice — to exactly compute the minimum-cost cut(s), despite the apparent intractability of the optimization problem (Cormen, Leiserson, and Rivest, 1990; Ahuja, Magnanti, and Orlin, 1993).\(^2\)

In contrast, other graph-partitioning problems that have been previously used to formulate NLP classification problems\(^3\) are NP-complete (Hatzivassiloglou and McKeown, 1997; Agrawal et al., 2003; Joachims, 2003).

### 3 Evaluation Framework

Our experiments involve classifying movie reviews as either positive or negative, an appealing task for several reasons. First, as mentioned in the introduction, providing polarity information about reviews is a useful service: witness the popularity of www.rottentomatoes.com. Second, movie reviews are apparently harder to classify than reviews of other products (Turney, 2002; Dave, Lawrence, and Pennock, 2003). Third, the correct label can be extracted automatically from rating information (e.g., number of stars). Our data\(^4\) contains 1000 positive and 1000 negative reviews all written before 2002, with a cap of 20 reviews per author (312 authors total) per category. We refer to this corpus as the **polarity dataset**.

**Default polarity classifiers** We tested support vector machines (SVMs) and Naïve Bayes (NB). Following Pang et al. (2002), we use unigram-presence features: the \( i \)th coordinate of a feature vector is 1 if the corresponding unigram occurs in the input text, 0 otherwise. (For SVMs, the feature vectors are length-normalized). Each default document-level polarity classifier is trained and tested on the extracts formed by applying one of the sentence-level subjectivity detectors to reviews in the polarity dataset.

**Subjectivity dataset** To train our detectors, we need a collection of labeled sentences. Riloff and

---


\(^3\)Graph-based approaches to general clustering problems are too numerous to mention here.

\(^4\)Available at www.cs.cornell.edu/people/pabo/movie-review-data/ (review corpus version 2.0).
Wiebe (2003) state that “It is [very hard] to obtain collections of individual sentences that can be easily identified as subjective or objective”; the polarity-dataset sentences, for example, have not been so annotated. Fortunately, we were able to mine the Web to create a large, automatically-labeled sentence corpus. To gather subjective sentences (or phrases), we collected 5000 movie-review snippets (e.g., “bold, imaginative, and impossible to resist”) from www.rottentomatoes.com. To obtain (mostly) objective data, we took 5000 sentences from plot summaries available from the Internet Movie Database (www.imdb.com). We only selected sentences or snippets at least ten words long and drawn from reviews or plot summaries of movies released post-2001, which prevents overlap with the polarity dataset.

**Subjectivity detectors** As noted above, we can use our default polarity classifiers as “basic” sentence-level subjectivity detectors (after retraining on the subjectivity dataset) to produce extracts of the original reviews. We also create a family of cut-based subjectivity detectors; these take as input the set of sentences appearing in a single document and determine the subjectivity status of all the sentences simultaneously using per-item and pairwise relationship information. Specifically, for a given document, we use the construction in Section 2.2 to build a graph wherein the source $s$ and sink $t$ correspond to the class of subjective and objective sentences, respectively, and each internal node $v_i$ corresponds to the document’s $i$th sentence $s_i$. We can set the individual scores $\text{ind}_1(s_i) \to P_{NB}^{\text{sub}}(s_i)$ and $\text{ind}_2(s_i) \to 1 - P_{NB}^{\text{sub}}(s_i)$, as shown in Figure 3, where $P_{NB}^{\text{sub}}(s)$ denotes Naive Bayes’ estimate of the probability that sentence $s$ is subjective; or, we can use the weights produced by the SVM classifier instead. If we set all the association scores to zero, then the minimum-cut classification of the sentences is the same as that of the basic subjectivity detector. Alternatively, we incorporate the degree of proximity between pairs of sentences, controlled by three parameters. The threshold $T$ specifies the maximum distance two sentences can be separated by and still be considered proximal. The non-increasing function $f(d)$ specifies how the influence of proximal sentences decays with respect to distance $d$; in our experiments, we tried $f(d) = 1, e^{1-d}$, and $1/d^2$. The constant $c$ controls the relative influence of the association scores: a larger $c$ makes the minimum-cut algorithm more loath to put proximal sentences in different classes. With these in hand, we set for $j > i$

$$\text{assoc}(s_i, s_j) \equiv \begin{cases} f(j - i) \cdot c & \text{if } (j - i) \leq T; \\ 0 & \text{otherwise.} \end{cases}$$

**4 Experimental Results**

Below, we report average accuracies computed by ten-fold cross-validation over the polarity dataset. Section 4.1 examines our basic subjectivity extraction algorithms, which are based on individual-sentence predictions alone. Section 4.2 evaluates the more sophisticated form of subjectivity extraction that incorporates context information via the minimum-cut paradigm.

As we will see, the use of subjectivity extraction can in the best case provide satisfying improvement in polarity classification, and otherwise can at least yield polarity-classification accuracies indistinguishable from employing the full review. At the same time, the extracts we create are both smaller on average than the original document and more effective as input to a default polarity classifier than the same-length counterparts produced by standard summarization tactics (e.g., first- or last-N sentences). We therefore conclude that subjectivity extraction produces effective summaries of document sentiment.

**4.1 Basic subjectivity extraction**

As noted in Section 3, both Naive Bayes and SVMs can be trained on our subjectivity dataset and then used as a basic subjectivity detector. The former has somewhat better average ten-fold cross-validation performance on the subjectivity dataset (92% vs. 90%), and so for space reasons, our initial discussions will focus on the results attained via NB subjectivity detection.

Employing Naive Bayes as a subjectivity detector ($\text{ExtractNB}$) in conjunction with a Naive Bayes document-level polarity classifier achieves 86.4%
accuracy. This is a clear improvement over the 82.8% that results when no extraction is applied (Full review); indeed, the difference is highly statistically significant ($p < 0.01$, paired t-test). With SVMs as the polarity classifier instead, the Full review performance rises to 87.15%, but comparison via the paired t-test reveals that this is statistically indistinguishable from the 86.4% that is achieved by running the SVM polarity classifier on Extract$_{NB}$ input. (More improvements to extraction performance are reported later in this section.)

These findings indicate that the extracts preserve (and, in the NB polarity-classifier case, apparently clarify) the sentiment information in the originating documents, and thus are good summaries from the polarity-classification point of view. Further support comes from a “flipping” experiment: if we give as input to the default polarity classifier an extract consisting of the sentences labeled objective, accuracy drops dramatically to 71% for NB and 67% for SVMs. This confirms our hypothesis that sentences discarded by the subjectivity extraction process are indeed much less indicative of sentiment polarity.

Moreover, the subjectivity extracts are much more compact than the original documents (an important feature for a summary to have): they contain on average only about 60% of the source reviews’ words. (This word preservation rate is plotted along the x-axis in the graphs in Figure 5.) This prompts us to study how much reduction of the original documents subjectivity detectors can perform and still accurately represent the texts’ sentiment information.

We can create subjectivity extracts of varying lengths by taking just the N most subjective sentences from the originating review. As one baseline to compare against, we take the canonical summarization standard of extracting the first N sentences — in general settings, authors often begin documents with an overview. We also consider the last N sentences: in many documents, concluding material may be a good summary, and www.rottentomatoes.com tends to select “snippets” from the end of movie reviews (Beineke et al., 2004). Finally, as a sanity check, we include results from the N least subjective sentences according to Naive Bayes.

Figure 4 shows the polarity classifier results as N ranges between 1 and 40. Our first observation is that the NB detector provides very good “bang for the buck”: with subjectivity extracts containing as few as 15 sentences, accuracy is quite close to what one gets if the entire review is used. In fact, for the NB polarity classifier, just using the 5 most subjective sentences is almost as informative as the Full review while containing on average only about 22% of the source reviews’ words.

Also, it so happens that at N = 30, performance is actually slightly better than (but statistically indistinguishable from) Full review even when the SVM default polarity classifier is used (87.2% vs. 87.15%). This suggests potentially effective extraction alternatives other than using a fixed probability threshold (which resulted in the lower accuracy of 86.4% reported above).

Furthermore, we see in Figure 4 that the N most-subjective sentences account for a sizeable majority of the source review’s words: The average word preservation rate is 66% when only one sentence is extracted, rising to 89% when 30 sentences are used. The average reduction in number of sentences is also dramatic: the average number of sentences in the source reviews is 32.3, while the average number of sentences in the subjectivity extracts is 3.5. Note also that these results are highly robust to changes in the polarity-classification threshold: for N = 30, the average accuracy drops by at most 1.5% when the threshold is increased to 0.75.

These findings indicate that the extracts preserve (and, in the NB polarity-classifier case, apparently clarify) the sentiment information in the originating documents, and thus are good summaries from the polarity-classification point of view. Further support comes from a “flipping” experiment: if we give as input to the default polarity classifier an extract consisting of the sentences labeled objective, accuracy drops dramatically to 71% for NB and 67% for SVMs. This confirms our hypothesis that sentences discarded by the subjectivity extraction process are indeed much less indicative of sentiment polarity.

Moreover, the subjectivity extracts are much more compact than the original documents (an important feature for a summary to have): they contain on average only about 60% of the source reviews’ words. (This word preservation rate is plotted along the x-axis in the graphs in Figure 5.) This prompts us to study how much reduction of the original documents subjectivity detectors can perform and still accurately represent the texts’ sentiment information.

We can create subjectivity extracts of varying lengths by taking just the N most subjective sentences from the originating review. As one baseline to compare against, we take the canonical summarization standard of extracting the first N sentences — in general settings, authors often begin documents with an overview. We also consider the last N sentences: in many documents, concluding material may be a good summary, and www.rottentomatoes.com tends to select “snippets” from the end of movie reviews (Beineke et al., 2004). Finally, as a sanity check, we include results from the N least subjective sentences according to Naive Bayes.

Figure 4 shows the polarity classifier results as N ranges between 1 and 40. Our first observation is that the NB detector provides very good “bang for the buck”: with subjectivity extracts containing as few as 15 sentences, accuracy is quite close to what one gets if the entire review is used. In fact, for the NB polarity classifier, just using the 5 most subjective sentences is almost as informative as the Full review while containing on average only about 22% of the source reviews’ words.

Also, it so happens that at N = 30, performance is actually slightly better than (but statistically indistinguishable from) Full review even when the SVM default polarity classifier is used (87.2% vs. 87.15%). This suggests potentially effective extraction alternatives other than using a fixed probability threshold (which resulted in the lower accuracy of 86.4% reported above).

Furthermore, we see in Figure 4 that the N most-subjective sentences account for a sizeable majority of the source review’s words: The average word preservation rate is 66% when only one sentence is extracted, rising to 89% when 30 sentences are used. The average number of sentences in the source reviews is 32.3, while the average number of sentences in the subjectivity extracts is 3.5. Note also that these results are highly robust to changes in the polarity-classification threshold: for N = 30, the average accuracy drops by at most 1.5% when the threshold is increased to 0.75.

---

9 This result and others are depicted in Figure 5; for now, consider only the y-axis in those plots.

10 Recall that direct evidence is not available because the polarity dataset’s sentences lack subjectivity labels.
subjective-sentences method generally outperforms the other baseline summarization methods (which perhaps suggests that sentiment summarization cannot be treated the same as topic-based summarization, although this conjecture would need to be verified on other domains and data). It’s also interesting to observe how much better the last $N$ sentences are than the first $N$ sentences; this may reflect a (hardly surprising) tendency for movie-review authors to place plot descriptions at the beginning rather than the end of the text and conclude with overtly opinionated statements.

### 4.2 Incorporating context information

The previous section demonstrated the value of subjectivity detection. We now examine whether context information, particularly regarding sentence proximity, can further improve subjectivity extraction. As discussed in Section 2.2 and 3, contextual constraints are easily incorporated via the minimum-cut formalism but are not natural inputs for standard Naive Bayes and SVMs.

Figure 5 shows the effect of adding in proximity information. $\text{Extract}_{NB+\text{Prox}}$ and $\text{Extract}_{SVM+\text{Prox}}$ are the graph-based subjectivity detectors using Naive Bayes and SVMs, respectively, for the individual scores; we depict the best performance achieved by a single setting of the three proximity-related edge-weight parameters over all ten data folds$^{13}$ (parameter selection was not a focus of the current work). The two comparisons we are most interested in are $\text{Extract}_{NB+\text{Prox}}$ versus $\text{Extract}_{NB}$ and $\text{Extract}_{SVM+\text{Prox}}$ versus $\text{Extract}_{SVM}$.

We see that the context-aware graph-based subjectivity detectors tend to create extracts that are

---

$^{13}$Parameters are chosen from $T \in \{1, 2, 3\}$, $f(d) \in \{1, e^{1-d}, 1/d^2\}$, and $c \in [0, 1]$ at intervals of 0.1.
more informative (statistically significant so (paired t-test) for SVM subjectivity detectors only), although these extracts are longer than their context-blind counterparts. We note that the performance enhancements cannot be attributed entirely to the mere inclusion of more sentences regardless of whether they are subjective or not — one counter-argument is that Full review yielded substantially worse results for the NB default polarity classifier — and at any rate, the graph-derived extracts are still substantially more concise than the full texts.

Now, while incorporating a bias for assigning nearby sentences to the same category into NB and SVM subjectivity detectors seems to require some non-obvious feature engineering, we also wish to investigate whether our graph-based paradigm makes better use of contextual constraints that can be (more or less) easily encoded into the input of standard classifiers. For illustrative purposes, we consider paragraph-boundary information, looking only at SVM subjectivity detection for simplicity’s sake.

It seems intuitively plausible that paragraph boundaries (an approximation to discourse boundaries) loosen coherence constraints between nearby sentences. To capture this notion for minimum-cut-based classification, we can simply reduce the association scores for all pairs of sentences that occur in different paragraphs by multiplying them by a cross-paragraph-boundary weight $w \in [0, 1]$. For standard classifiers, we can employ the trick of having the detector treat paragraphs, rather than sentences, as the basic unit to be labeled. This enables the standard classifier to utilize coherence between sentences in the same paragraph; on the other hand, it also (probably unavoidably) poses a hard constraint that all of a paragraph’s sentences get the same label, which increases noise sensitivity.\footnote{For example, in the data we used, boundaries may have been missed due to malformed html.}

Our experiments reveal the graph-cut formulation to be the better approach: for both default polarity classifiers (NB and SVM), some choice of parameters (including $w$) for Extract\textsc{SVM+Prox} yields statistically significant improvement over its paragraph-unit non-graph counterpart (NB: 86.4\% vs. 85.2\%; SVM: 86.15\% vs. 85.45\%).

5 Conclusions

We examined the relation between subjectivity detection and polarity classification, showing that subjectivity detection can compress reviews into much shorter extracts that still retain polarity information at a level comparable to that of the full review. In fact, for the Naive Bayes polarity classifier, the subjectivity extracts are shown to be more effective input than the originating document, which suggests that they are not only shorter, but also “cleaner” representations of the intended polarity.

We have also shown that employing the minimum-cut framework results in the development of efficient algorithms for sentiment analysis. Utilizing contextual information via this framework can lead to statistically significant improvement in polarity-classification accuracy. Directions for future research include developing parameter-selection techniques, incorporating other sources of contextual cues besides sentence proximity, and investigating other means for modeling such information.

Acknowledgments

We thank Eric Breck, Claire Cardie, Rich Caruana, Yejin Choi, Shimon Edelman, Thorsten Joachims, Jon Kleinberg, Oren Kurland, Art Munson, Vincent Ng, Fernando Pereira, Ves Stoyanov, Ramin Zabih, and the anonymous reviewers for helpful comments. This paper is based upon work supported in part by the National Science Foundation under grants ITR/IM IIS-0081334 and IIS-0329064, a Cornell Graduate Fellowship in Cognitive Studies, and by an Alfred P. Sloan Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the National Science Foundation or Sloan Foundation.

References


Qu, Yan, James Shanahan, and Janyce Wiebe, editors. 2004. *AAAI Spring Symposium on Explor-